Knowledge Organisation in Historical Information Systems

Javier García-Marco, University of Zaragoza, Spain

Abstract: The complexity and diversity of historical information makes the field of Historical Information Systems (HIS) an interesting platform for Knowledge Organisation research. HIS are defined considering their specific aims and their role as a knowledge store and transfer system. Improving such a system requires the development of representational strategies, solutions for common retrieval problems and a classification capable to organise it all. Some suggestions are made in these directions.

Historical Information Systems (HIS): the concept

A historical information system (HIS) may be defined as an automated system integrating a set of databases and formal procedures, designed and maintained to store, treat and retrieve historical information. They must be able to store sources —both references and reproductions—, bibliographical references and research work; should they be textual, graphic or procedural. The stress is put in the interface among the different types of information. The HIS is considered to be an open evolutionary system, growing towards an ever closer integration of the data. Finally, the HIS must be considered as a part of a scientific network of research and custodial centres, with which it must interchange data and knowledge, and therefore pursue co-operative normalisation policies.

Despite the growing research in special information systems, e. g. geographical, business, management, archival and library information systems, HIS are advancing at a lower pace. Perhaps, this is due to the usually individualistic work of historians, who seldom consider giving access to their data to other scientists or doing co-operative research. This is usually so, unless they are tied by very strong institutions, or unless the accumulation of electronic records is the result of the activities of the original organisation. However, historical information is useful for many sciences and activities. Besides it, the field of HIS is nowadays divided among historical archives and archival networks, heritage databases and historical research projects. Advances in these different areas will undoubtedly affect the others, but the fact is that there is insufficient contact among them. This gap is the result of different stresses: historical researches pursue creation of knowledge from historical sources; heritage and archival officers look mainly for control and preservation of the documentary units in which sources are stored.

Anyway, historical work is rapidly changing: it is leaving the craftsmanship stage, and there is a strong trend towards data-processing and database developing; and,

as the need grows, towards data-interchange among the new data banks\textsuperscript{2n}. In fact, the development of interconnected historical information systems will change the way in which much of the historical research is being done. This will produce important profits. First, much work that is being redone by each historian approaching the same historical sources will be accessible to others. Second, this trend will tighten the evolution towards a more "normal" historical science, in the sense that was casted by Thomas Kuhn. More and more historical hypothesis generation and testing will be done in terms of formal procedures implemented over data-structures. Improvement of data and peer-checking of hypothesis will be enhanced. Third, institutions will appear that will take the responsibility of the collection, maintenance and dissemination of this information and of the development of group-research policies, ensuring therefore a less dispersed historical scientific accumulation. Lastly, this process is currently being enhanced by data interchange among research institutions.

The advance in information-based historical research will also result in new bridges between researchers and conservationists. As HIS enlarge, database and research managers will feel a stronger need for information control. As HIS interchange more and more information, there will be a powerful bias towards representational standards, which are more evolved in the archival and custodial world (International Council on Archives, 1992, 1993; Cook & Procter, 1989; Cook, 1993).

\section*{The knowledge-transfer process}

Theoretically, a HIS may be represented as a special case of an automated knowledge transfer system (García Marco, 1994). It should be able to help in the acquisition, representation and transfer of historical knowledge. As in any other science, historical hypotheses are the result of identification, classification and inter-relation (throughout predication) of historical facts. Testing of hypothesis is made by ensuring: a) the assumption that data correspond to 'real' facts — throughout their internal, external and statistical criticism —; b) that the methodology drives cleanly to the results and conclusions; and c) that the theoretical model is a parsimonious and coherent representation of the reality, which is established through derivation of new hypotheses —implications and generalisations— and their contrasting with well-founded historical facts. As societies, institutions and groups are now overwhelmingly considered to be systems where their different causal variables —social, economic, political, cultural, religious, etc.— are related, historical research is oriented toward the establishing of such relations, and cross-testing of hypothesis results in this way facilitated.
Primary documents  
<table>
<thead>
<tr>
<th>Historical Sources</th>
<th>Documentary representations</th>
<th>Information needs and answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Historical papers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free text retrieval</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1. Knowledge representation process in a HIS**

Historical work grows as an interaction of three different forces: a) the historical source; b) the Historiography, which are historical interpretations made by previous historians; and c) the work of the actual historian, with its background, theories, knowledge of the sources and methodology. All of them are a product or a subsystem of the whole social system in which they evolve. The system should represent all these factors and pursue the integration among them. Finally, the system must be able to facilitate the transference of the historical knowledge, expressed throughout terms and their relations, throughout the research community and among its different groups. This is—as in any other process of communication—a question of building a common code, both in its vocabulary and grammar; or, better expressed considering its artificial genesis, a normalisation process (see note 2 for bibliographical references).

A graphical sketch of a HIS is offered downwards. A HIS is submitted as a representational system of primary documents—both sources and Historiography—guided by the informational needs of the users, and the existing theories about historical development, knowledge representation and information management. The document representations act as a bridge between the informational content of the primary documents and the users' informational needs and cognitive representations of reality, and among their correlative linguistic expressions.

**Representational strategies**

We have tried to defend the need of reflection about HIS and to draft their basic theoretical scheme. However, although the existence of a gap, a need and an op-
Portunity is clear, their specific aims, basic structure and representational approaches are to be sketched.

1. **Structural representation.** The structure of the system should be able to deal with and inter-relate image, documentary and relational databases, combining know-how from, at least, archivists, information scientists, historians and language experts. We are thinking of systems able to store, inter-relate and manage optical reproductions, full-text transcriptions, abstracts, indexing and classification data; together with archival data, bibliographic references, research project’s management information, raw data obtained from documents for specific scientific purposes, and results of these analyses (document reconstruction, statistical results and graphics).

![Diagram of levels of representation and their organisation](image)

Fig. 2. Levels of representation and their organisation

The system will result of natural accumulation of research. They should ensure reuse of, improvement of and navigation through any kind of document representation, and accumulation and checking of historical data. Some sort of standard for historical referencing would even ensure the hyper-textual connection of the information systems with the scientific papers produced by researchers who use them.

2. **Dynamic representation.** The relations among the concepts contained in the primary documents can be expressed in two ways, depending on the kind of their original structure. Two kinds of textual structures in primitive documents can be distinguished: a) very fixed structures, with a rigid textual structure where only some variable elements change, like most of notarial and fiscal documentation; and b) very rich and complex textual structures, usually narrative ones, which allow different lectures and interpretations. The former may be represented in quite
rigid, relational data structures, and allow quicker, more exhaustive and precise retrieval. The later require long and very diverse analytical procedures; they are best accessed in a first stage through text retrieval approaches, included hypertext; they generate many analytical data sub-products, like lexical and statistical data; and produce very complex relations with the Historiography.

**Retrieval problems and possible solutions**

1. *Retrieval problems.* Each one of the different representational levels that we have considered produces specific knowledge access problems, and only a connection among the different representational levels will produce a true and complete knowledge organisation in the HIS. In fact, vocabulary control is critical considering the nature of historical information, affected by lexical, phonetic and orthographic variations, in the source level, and of terrible terminological problems in the level of academic and scientific discourse.

2. *Top-to-down and down-to-top strategies.* Both top-to-down and down-to-top strategies will be useful: the first ones proceeding from the pragmatic and semantic level down to the significant level, ensuring the building of terminological systems; and the second ones putting order in the significant level and connecting it with the meaning level. As a result of the former approach, indexing vocabulary (not only generic subjects, but topographic and personal names too) will be controlled directly in abstracts and indexed fields and properly organised in a documentary language. This will ensure maximum exhaustivity and precision in retrieval. On the other side, parsing algorithms should be also developed to aid search in transcription-resulting free-text — usually the result of different languages, dialects and graphical conventions —, facilitating the identification of graphical variations and synonyms.

3. *The need of an interface between the emic and etic vocabulary.* The conceptual level of a HIS is affected by the divergence between the original vocabulary and organisation extracted from the documents (emic vocabulary), on the one hand, and the conceptual organisation and vocabulary used by the researchers to describe the historical society (etic vocabulary), on the other. Historical work progresses usually from the original concepts as they are expressed in the vocabulary of the epoch, which are classified and switched to the prevailing theories on the evolution of society/culture and/or their specific facets. The vocabularies of the different theories have their own implications and relations, and their terms have not necessarily the same meaning when interchanged, so there is also a need of interfacing them (both connecting synonyms and distinguishing homonyms). Therefore, it is a central aspect of a HIS the inter-relation of the different conceptual systems.
4. **Adding phonetic knowledge for free text retrieval.** The representation of historical documents in retrieval systems is at least as expensive as common bibliographic representation, and in some aspects—for example when complex summarising is needed—far more expensive. At the same time, optical character recognition (OCR) is rapidly improving, and it is beginning to be commercially viable for handwriting. This may indicate that we are in the dawn of the application of these technologies to the historical sources, and therefore, of getting a cheap free-text access to the documents. But this kind of sources has a peculiarity that is well known to all that work with them, and that grows as the texts are older; that is: phonetic and spelling variations. The older the text is, the less linguistic normalisation exists, and the bigger the vocabulary control problems are. The use of the usual free-text facilities—truncation, masking and proximity operators—will produce too much noise. The two only approaches left to solve these problems are: a) lexical treatment through a build-in thesaurus; and b) procedural or indexed representation of phonetic and spelling knowledge. The results of the second approach will always be partial, but these capabilities would improve dramatically free-text search performance, and would not be too costly. Along with it, similar strategies have been developed to deal with spelling errors when searching in modern bibliographic databases (Walker & Jones, 1987). Basic linguistic processing, like stemming, will provide with a somewhat hypertext-like navigation utility for complex narrative documents.

5. **Adding intelligent procedures.** Data-processing scripts for hypothesis generation and testing must be made explicit and documented to ensure future checking and re-using. Besides it, the HIS may be enriched with procedural expert-system-like knowledge representation in quite many specialised fields, for example: reference work, through the automation of research guides to historical sources; or complex assisted search through different facets in a changing administrative and social environment (Esteban & García, 1992).

**A sketch of the conceptual system**

The conceptual system should be a faceted one to ensure access from different points of view and to avoid inconsistencies not very practical for on-line retrieval—such as the treatment of national history in class 9 of the UDC in contrast to the use of common place auxiliaries—. The faceted systems, as originally designed by Ranganathan, are also the best suited to solve the typical problem of fluid sub-discipline divisions and merges that occur in History and in the rest of the other social sciences. The scheme should be able to deal with both sources, research and teaching work. These are the facets we are considering at the moment:

a) **Historical singulars**: physical and juridical persons, and their relations. A very serious problem with this facet is the personifying or references to personality
entities through spatial entities (for example, the role of Spain in the Second World War); perhaps we should consider to leave the spatial facet only for natural places, and distinguish the use of political institutions as places or personalities by means of a role operator when indexing or classifying.

b) **Historiographical singulars** (only for historiographic documents): this facet could absorb the authority file of authors and institutions (research institutions, libraries, archives, museums) and codify the relations among them. This facet and the former could be merged in an only personality one, distinguishing both subfacets through a role operator both in the authority file —if their sense is unique— and when indexing.

c) **Objects and artefacts**.

d) **Specific scientific categories and constructs** applied to the other facets.

e) **Action**, which could be expressed by extracted terms as the document diplomatic or juridical type, concrete terms, like war or revolt, or assigned constructs, such as progress, decadence, revolution, etc.

f) **Great themes** or ‘social institutions’: society, economy, religion, culture, arts, policy, military and war, science, technology, etc.

g) **Disciplinary subdivisions**: This facet may be sometimes redundant in relation with the ‘great themes’ facet, because it is usually build from them (for example, economic history), but it would ensure rapid disciplinary access to the HIS. We must have in account that the relation between themes and discipline is always evolving and it is not bi-univocal: disciplines depend on the sociological organisation of science and take in their scope different subjects and relations among them.

Papers on history of the economy can be written by scholars not necessarily ranged in the field of the Economic History.

h) **Spatial context**, with its subdivisions. As the social space is not homogeneous, but fragmented and multidimensional, this facet should be sub-faceted or cross-faceted by social institutions and themes (political, religious, etc.)

i) **Temporal context**, reducing events to our relative notation, although an authority file in the form of a calendar could be quite useful. The UDC provides a quite good solution for temporal notations. A problem to be considered is the non temporal connotations of temporal environment descriptors: v. g. French Revolution is not equal to France + 1789.

j) **Theoretical systems and paradigms** (for historical literature only).

k) **Methodology** (for academical work only).
I) Form of the document (monograph, handbook, dissertation, computer file, facsimile, diplomatic type, etc.).  
The system should provide a fixed notation order and the automatic retrieval system should ensure the indexing of each conceptual element of the classification number. The HIS should provide build-in and/or assisted automatic broadening and narrowing search procedures, in order to allow maximum precision.

Conclusion

The HIS research area, although it is not a critical economic one, represents a privileged testing field for many theoretical questions in knowledge organisation and representation research, because of its complexity. Along with it, HIS might be a path to overcome the conflicts existing between archival representational procedures—based in the preservation of the original order, which, on the other hand, ensures some kind of 'knowledge archaeology'—and the informational needs of the users, and a way of disciplining historical research, and therefore, ensuring peer-to-peer communication and cross-testing of hypothesis and theories among historians.

Bibliography


Notes

1. The subject of applications of computer science to historical research is a well established sub-discipline of the historical studies, based on good studies and handbooks (for example and recent, Reiff, 1991; Mawdsley & Munck, 1993), scientific conferences (CNRS, 1980; see note 3), and a scientific association specially devoted to this subject: the "Association
for History and Computing", which publishes the periodical *History and Computing* (1989-), promotes meetings (Denley, 1987; Denely et al., 1989; Mawdsley, 1990) and has published a guide for historical data files in machine readable form (Schurer & Anderson, 1992)—. A subject of growing importance is the application of computers to the teaching of History (Schick, 1990).

2. The standardisation of historical databases is nothing new and has been considered in a number of pioneering scientific meetings (Fossier, 1977; Genet, 1988; Hausmann, 1987; Thaller, 1986).

3. The distinction between emic and etic aspects of the culture was established by the anthropologist Kenneth Pike, and developed by Marvin Harris (1968). The emic aspect of the culture refers to the semantic and communicative phenomena of the natives and to the their inner states and feelings. The etic aspect refers to the concepts and propositions that are relevant to the scientific community that studies the culture.

4. An interesting facility would be a syndetic device to connect descriptors that represent jointly an interdisciplinary or specialised field through their interconnected subjects, in contrast to those denoting the same themes connected empirically. This is, for example, an item may have the descriptors 'society' and 'rural' because it works with the problem of the "The rural communities of Aragón (Spain) in the Middle Ages", and another because it comes specifically from the field of peasant studies. The relation between the strongly connected descriptors and the specific historical sub-disciplines—which change in the different historiographic traditions—could be codified through an authority file. This approach would provide an alternative solution for the description of the disciplinary point of view to that of using specific descriptors.